Disease Prediction using Machine Learning Techniques in Healthcare

Rashmi V. Shinde

Department of Computer Engineering, Gokhale Education Society's
R. H. Sapat College of Engineering Management Studies and Research, Nashik-05

***Abstract*: In recent days big data is one of the fastest and widely used approach in each and every field. By taking the help of huge amount of data biomedical and health care areas reaches their progress and also this huge amount of data profit a perfect medical data investigation, quick disease forecasting, correct data about patient can be confidentially stored and used for predicting the disease. Furthermore the correctness of an analysis can be reduced because the number of reason like imperfect medical data, some area wise disease features which can be outbreaks the prediction. In this paper we can use a various machine learning based approach for the correct disease prediction for such prediction we can gather the hospital related data of a specific area. For imperfect data the Stochastic gradient decent method is use to accomplish the incompleteness of data. For predicting disease, in the earlier days Unimodal Disease Risk Prediction approach of CNN (CNN-UDRP) is applicable. But there are some limitation for CNN-UDRP as it consider only labelled or structure data so to overcome the limitation of CNN-UDRP approach we concentrate on other CNN-MDRP approach as it works on both labeled and unlabeled type of data. Still now the existing systems are not feasible for working with different type of data that's why the CNN-MDRP approach is more suitable for predicting the diseases with respect to other approaches.**

***Index Terms*: Big data analytics, Machine Learning, Disease prediction, Healthcare.**

I. INTRODUCTION

In any domain we obtain the result in terms of big data, data can be called as big data if application produce the more volume of data from their processes. The big data have their own important three properties that are volume mean number of data, variety mean various type of data and velocity mean processing speed of data. As we know the data which is produced from medical sector is in very large volume and we called it as one of the best application of big data. The hospital facilities are available in variety of areas and each healthcare have information about their patient it mean the data produced from healthcare is from number areas and this data is taken together for predicting the disease. For guessing the diseases we consider of some Machine Learning based approaches because the approaches which are based on machine learning are liable to display the more accurate outcome of prediction.

In position with forecasting the disease several existing learning is focused only on well arranged (structured) data. And for the free (unstructured) data make use of different approach in Machine Learning such as Convolutional Neural Network (CNN). CNN Neurons are used to make Neural Network, in this process the input is given to each and every neuron after that neuron perform some procedure on it and form a network. For accurate prediction of the disease firstly we have efficient data for processing, we got the efficient information by reducing incompleteness of particular set of data. As the medical outcome is very important to every one so it is necessary to provide accurate result of patient and for providing the accurate result we use some classification technique like Naive bayes classification it is used to classify the similar type of data into single group. Another classifier is K-nearest neighbor which gives the k number of closest value from particular dataset along as per the particular input data.

II. REVIEW OF LITERATURE

Shah A.et al. [2] here writer planned the approximation analytics. In mention method the number of data sources is essential for gathering the informational which needed for prediction. And this is likewise responsible to accumulate the crucial data by multiple things such as mobile phones and some presentations with respect to general things. This information is combined to determination the clinical approximation analysis. Writer gather related information for the purpose of health risk prediction. While going through this process many problems are generated. As per the risk of particular patient necessary aspects are used like to improve the goodness EHR is used here.

B. Lee et al. [3] in this paper writer offered method for measuring the score of heart disease. There are various indicators on which the score of heart disease is depend. Only measuring the heart score have some issue that it is not liable to decide whether the patient is safe or unsafe for discharge. With help of previous record of patient they are able to determine risk about heart disease. For measuring the heart score ECG is most common aspect is come in picture.

In this paper V. Chaurasia et al. [4] plan an approach which is essential to estimating the heart related disease. They basically make use of various machine learning approaches. The first approach which used is decision tree. Along with decision tree approach also regression and classification approaches are considered. For taking proper decision, Decision tree is liable. For providing the accurate result in predicting diseases both classification approaches are implemented i.e. regression approach and second one is classification.

A. Apte et al. [5] here author define the fact which is essential for ill people who are suffering from heart attacks. For classifying and preprocessing the data one of the classifier is used called as naïve bayes classifier. Also various data mining approaches are considered for testing over prediction, the approaches are Decision tree another one i.e. naïve Bayesian and KNN. As all these approaches are apply for estimating the chances of heart attack for patient.

Fellow C. et al. **[**6**]** provide smart dressing idea by taking help of several tools and technique. In this case multiple human signatures are collected through smart dressing and that human signs are related with human hobbies. Smart dressing includes movable fiber cables, electrode cables to generate ECG indicators.by collecting the various signatures it is possible to determine the body analysis for predicting the diseases.

Jihe Wang et al. [7] proposed the telehealth approach using discrete data format. In which BSN component is included in system, it has number of nodes into its network model so we are required to find out the bandwidth within multiple nodes. At each iteration the bandwidth between the nodes is measured and varies. BSN are responsible for collecting the signals from human body. According to the signal information from human body the experts are able to inform the patient about their chances of illness.

B. Guo et al. [8] proposed a smart health system which support to both techniques that is big data and cloud. In the of health CPS framework, three layers are available for gathering data, managing data and display data. In first layer data can be gather from customer, clinical area, internet all collected data then pass towards next layer management layer which have distributed file storage and distributed cloud computing for managing the data. After that the management layer interact with view layer in which the user interact with the application through interface of user, program interface and multiple data access tools.

In this paper, H. Li. et al. [9] planned for design a system for patient who are go through with certain diseases, author proposed an active method for identifying the similar symptoms of different patients. They check that symptoms of patient A or patient B is match with patient C for this purpose they use the Active Risk Prediction approach. In ARP approach it take input of EHRs and another input is patient's symptoms and pass for training in that the medical expert find out the similarities between patients. According to the result from ARP suggest to patient about the disease they are suffering from.

In this paper S. Roy. et al. [10] define that the mentioning the patient readmission risk will improve the care about ill person. Day by day the number people are suffering from the heart failure so it is important to tell them about readmission and it is possible by classification of hierarchy dynamically. Many of the hospitals use this dynamic classification for patient readmission, this dynamic classification use the patient past data as well as current data for making the hierarchy. The readmission is depends on performance of patient within 30 days.

J. C. Lo. et al. [11] Define the network selection approach based on bio inspired mechanism. The bio inspired mechanism is defined mathematically by ASM Attractor selection system. Here they use the advance version of ASM i.e. EASM for taking the decision about particular network selection. EASM have self-adaptive mechanisms that allow many patients to randomly select the network in collective way. To give the qualitative service over the vehicular network the self-adaptive aspect for network selection is considered.

III. SYSTEM OVERVIEW

A. Problem Statement:

To define an effective and early resolution for predicting disease with the help of machine learning-based approaches works on both labelled and unlabeled data.

B. System Architecture:

The proposed systems initially take the massive data from dataset of particular disease as the system works on medical data dataset contain the bulky or massive data for particular disease. The data come from dataset is then consider for training. Sometime the dataset is imperfect or inappropriate which results in less accuracy of prediction, in such case it is necessary to complete the dataset it is possible by data imputation concept. With data imputation latent factor is used to take the appropriate data when there is missing data in the dataset. Dataset contain all type of information it can be structure data or free data, for sending data to testing first we need to represent it in structured and unstructured format mean the textual data is represented in well format. For better prediction classification is takes place, there are various classification algorithms are available for performing the classification on data present in dataset. Once the data is classified it will liable to take for further use. The tested data then given to the next phase training phase. In training phase the number of iteration are performed on dataset to train it by forming the neural network. Neural network is able test the conformation by itself which is nothing but the cross validation. In cross validation data is represented into two parts training and testing. CNN classifiers take input from tested data and perform classification on that data.

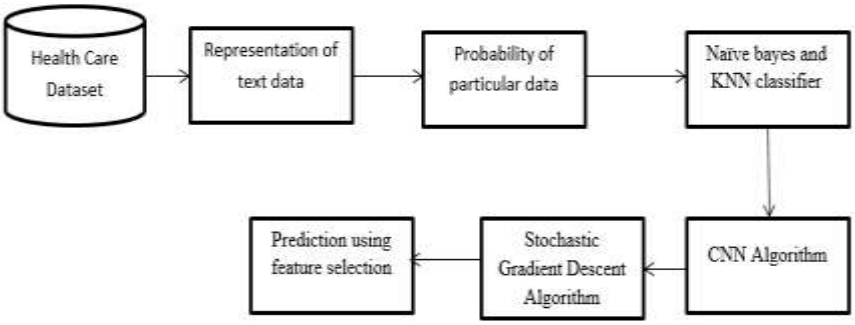

Fig.1 System Architechture

In next phase Stochastic gradient decent algorithm is considered to finding out the error rate. In error rate calculation we check the performance of data module over past recognized data. In case the error rate is max then we again assign the different weight to neuron and perform the operation again it will continue up to we got the minimum error rate. For taking the textual characteristics automatically CNN is best over other approaches. With basic CNN algorithm CNN-MDRP approach is used because it works on both labelled and unlabeled data. CNN-MDRP increases the prediction possibilities over beforehand selected characteristics. By using the feature selection approach some features are selected for predicting the chances of disease.

C. Algorithms:

1. Role of CNN MDRP in Hospitality:

According to previous study unimodal risk prediction approach works only on textual structured data for identifying risk of patient with particular disease. Another approach we can say that the advance version of UDRP is MDRP approach works any type of data including labeled, unlabeled and textual data. For prediction these approach take some features from text data and structure data. The CNN have multiple layer input layer, output layer and convolution layer, we pass input data to input layer from dataset. In next layer the weights and biases are assigning to the each neuron for processing in convolution layer. The weight is varying in each iteration to minimize the error rate and get desire output.

Step 1: Take input parameter from dataset.
Step 2: Input pass to convolution layer for processing.
Step 3: Convolution layer assign weight to each input neuron for further processing as w0, w1, w2 and so on.
Step 4: Get the output at output layer from convolution layer
Step 5: If the prediction have less accuracy reassign the weight to neuron and again perform processing.

2. Stochastic Gradient Descent Approach:

In machine learning domain SGD approach is commonly used for providing training to the data model. Neural network a Machine learning Algorithm have some constraint like weight and biases to identify the correctness of specific set of constraint. This is iterative approach in which we start its evolution by taking some input value to our model. At initial step if the output is inappropriate then we improve it slowly in next iteration by changing the values of weight and biases. Because of this the cost of given function is also minimize.

Step 1: Initialize the parameter and assign weight to them.
Step 2: Calculate the gradient according to parameter.
Step 3: Update the weight in case when we not get desired output.
Step 4: Repeat step 2 and 3 until we get appropriate output with minimum cost of particular function.

IV. RESULTS AND IMPLEMENTATION

Fig. 2 display percentage of risk prediction of disease according to input dataset. Fig. 3 here the feature selection technique is used to display the chances of disease prediction. Fig.4 display the result in graphical format related with the predicted disease. And Fig. 5 indicate the comparison of existing and proposed system.

.

Fig.2 CNN approach.

In CNN MDRP approach it displays the percentage of risk about perticular disease by considering the attributes in dataset. The attribute consider here are Admitted and Diabetic Med, the atrribute Admitted indicate the how many times the patient admitted in hospital and Dibetic Med indicate that the patient is having dibetic or not.

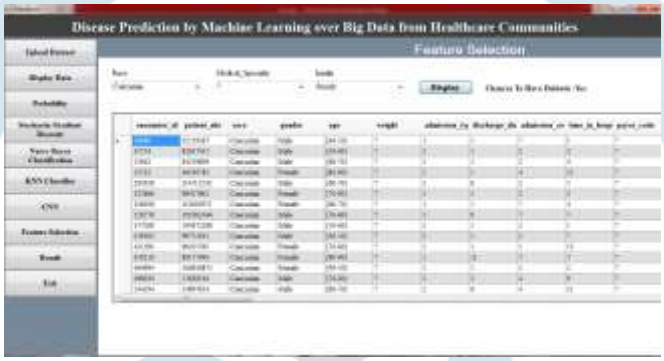

Fig.3 Feature Selection.

In feature selection technique we select the perticular attribute for further processing. The parameters selected here are Race, Medical Speciality and Insulin. These three attribute provide the more acurate prediction of chances of disease. The prediction is displays by considering the Diabetic Med and number of time the patient admitted in hospital.

Fig.4 Result.

According to result of feature selection graph is display the accuracy of that output along with the F1-measure. If the patient have specific medical speciality and insulin it tell the chnaces to have disease or not.

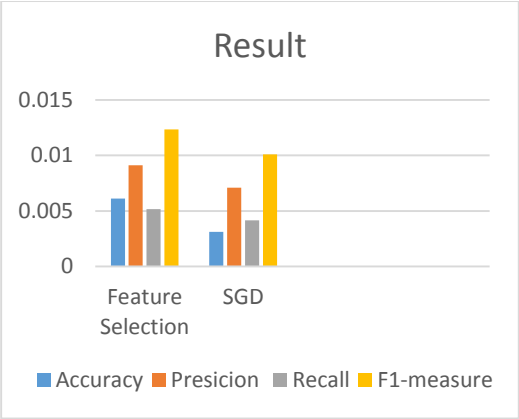

Fig. 5 Comaprision of result

In comparision of feature selection and SGD approach the result of festure selection is best over SGD approach. For this comparision we consider the three parameter for feature selection and five parameter for SGD from that three paramtera are common for both the approaches.

A. Dibetics Dataset:

Sr. No.	Dataset	No. of records	No. of attribute	Source
1	Diabetic data	101745	49	http://archive.ics.uci.edu/ml/machine-learning-databases/diabetes/

V.CONCLUSION

In this paper, we study about prediction of disease from large number of medical data. We propose feature selection technique for predicting the chances to have disease according to some selected features from dataset and a CNN-MDRP approach is used to predict the risk. The information produced from clinical area is in both structure and unstructured format. By using some machine learning approaches like CNN-UDRP which focuses only on well-organized data not a free data. But for high correctness in prediction we required to work on all type of data which can be either structured or free for that purpose the previously available approach under the machine learning concept is used called as MDRP approach based on CNN. So that the perfection in predicting the diseases is high in case of MDRP approach.

ACKNOWLEDGMENT

I am truthfully obligated and grateful to guide for their valuable guidance and inspiration. And also thankful to all staff members of Computer Department of GESRHSCOE-Nasik.

REFERENCES

[1] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", Electronic ISSN: 2169-3536.

[2] W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[3] S. Marcoon, A. M. Chang, B. Lee, R. Salhi, and J. E. Hollander, "Heart score to further risk stratify patients with low timi scores," *Critical pathways in cardiology*, vol. 12, no. 1, pp. 1–5, 2013.

[4] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques", Caribbean Journal of Science and Technology, vol.1, pp 208-217, 2013.

[5] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[6] M. Chen, Y. Ma, J. Song, C. Lai, B. Hu, "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring," ACM/Springer Mobile Networks and Applications, Vol. 21, No. 5, pp.825C845, 2016.

[7] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," *Journal of Systems Architecture*, vol. 72, pp. 69–79, 2017.

[8] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Healthcps: Healthcare cyber-physical system assisted by cloud and big data," *IEEE Systems Journal*, 2015.

[9] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 1070–1093, 2015.

[10] S. Basu Roy, A. Teredesai, K. Zolfaghar, R. Liu, D. Hazel, S. Newman,and A. Marinez, "Dynamic hierarchical classification for patient risk-ofreadmission, "in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2015, pp.1691–1700.

[11] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, "A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3033–3049, 2015.